Annals of Combinatorics 12 (2008) 45-51
0218-0006/08/010045-7
DOI 10.1007/s00026-008-0335-4

Annals of Combinatorics

A Subdivision Approach to Maximum Parsimony

Trevor C. Bruen[1] and David Bryant[2]

[1]Department of Mathematics, University of California, Berkeley, CA 94720, USA
tbruen@math.berkeley.edu

[2]Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand
d.bryant@auckland.ac.nz

Received January 16, 2006

AMS Subject Classification: 68R10, 68R05, 68Q25, 92D15

Abstract. Determining an optimal phylogenetic tree using maximum parsimony, also referred to as the Steiner tree problem in phylogenetics, is NP hard. Here we provide a new formulation for this problem which leads to an analytical and linear time solution when the dimensionality (sequence length, or number of characters) is at most two. This new formulation of the problem provides a direct link between the maximum parsimony problem and the maximum compatibility problem via the intersection graph. The solution for the "two character case" has numerous practical applications in phylogenetics, some of which are discussed.

Keywords: maximum parsimony, compatibility, character subdivision, Steiner tree

1. Introduction

Given a connected graph $G = (V, E)$, an edge weight $w(e) \in \mathbb{Z}_0^+$ for each $e \in E$ and a set of vertices $S \subseteq V$, the *Steiner tree* problem is to find a subtree $T = (V', E')$ of G such that $S \subseteq V'$ and the sum of all the edge weights is minimized ([8]). It is well known to be NP complete ([7]). A more restricted version of the general problem can be obtained by insisting the edge weights conform to some metric. For instance, consider a fixed alphabet A and the complete graph G on A^N (N is referred to here as the dimension) with edge weights defined as the Hamming distance d on A^N, i.e., $d((a_1, \ldots, a_n), (a'_1, \ldots, a'_n))$ is equal to the number of indices i such that $a_i \neq a'_i$. Then the phylogenetic Steiner tree problem (or maximum parsimony problem) is to find a Steiner tree for G whose vertex set includes $S \subseteq A^N$. This problem is also known to be NP complete ([6]).

For the most part, statistical methods for inferring phylogenies ([4, 11]) have supplanted maximum parsimony approaches in the construction of phylogenetic trees from conventional sequence data. Nevertheless, maximum parsimony is still widely used to

infer evolutionary trees based on morphological characters, to build supertrees, and to perform fast heuristic tree searches.

In this note we make two principal contributions:

i) an alternate formulation of the maximum parsimony problem in terms of subdivision; and
ii) detailed analysis of the two dimensional case (i.e., $N = 2$) showing that the problem can be solved not only in polynomial time, but actually in linear time.

The latter result is proved in two different ways and upper bounds on the maximum parsimony score for two characters are derived. The two characters results permit an efficient approach to detect genetic recombination ([12]), although there are potentially other applications such as improved lower bounds for parsimony (e.g., [3,5]). The result on two characters permits the efficient computation of the refined incompatibility score ([13, 15]) for two characters.

2. Notation and Definitions

Further details concerning mathematical phylogenetics and origin of the notation can be found elsewhere ([16]). Let X be a set of n species and χ be a function (called a *character*) from X to a finite set of states C. The number of states of χ $\big($cardinality of the image of $\chi\big)$ is denoted by $|\chi|$. Let $\pi(\chi)$ denote the partition of X induced by $\big\{\chi^{-1}(\alpha)\colon \alpha \in C\big\}$. Each equivalence class of $\pi(\chi)$ is referred to as a *block* of χ, with the number of blocks equal to $|\chi|$. A character χ' *refines* χ if every block of χ' is a subset of some block of χ, which holds if and only if $\chi'(u) = \chi'(v)$ implies $\chi(u) = \chi(v)$ for all $u, v \in X$. Note that the character with only one block is refined by all other characters, while the character with one block for each element in X refines all other characters. A *subdivision* of a character χ is the replacement of one block of the character with two disjoint and non-empty blocks.

An X-*tree* is an ordered pair $\mathcal{T} = (T, \phi)$ consisting of a tree T and a function $\phi\colon X \to V(T)$ with the property that every vertex of T of degree 1 or 2 is labeled by ϕ. A *phylogenetic* X-tree is an X-tree with the property that ϕ induces a bijection between X and the leaves of T . An *extension* of χ to an X-tree $\mathcal{T} = (T, \phi)$, is a function $\bar{\chi}$ from $V(T)$ to C such that the vertices of T are labeled in accordance with χ, i.e., $\bar{\chi} \circ \phi = \chi$.

Consider an extension $\bar{\chi}$ of some character χ to an X-tree $\mathcal{T}$ with underlying tree T. Then define $\mathrm{Ch}(\bar{\chi}, \mathcal{T}) := \big\{e = \{u, v\} \in E(T)\colon \bar{\chi}(u) \neq \bar{\chi}(v)\big\}$ and $\mathrm{ch}(\bar{\chi}, \mathcal{T}) := |\mathrm{Ch}(\bar{\chi}, \mathcal{T})|$. The *parsimony score* of χ on $\mathcal{T}$, $l_{\mathcal{T}}(\chi)$, is defined as the minimum of $\mathrm{ch}(\bar{\chi}, \mathcal{T})$ over all extensions of χ to $\mathcal{T}$. A character χ is *convex* on an X-tree $\mathcal{T}$ if and only if $l_{\mathcal{T}}(\chi) = |\chi| - 1$. For a sequence of k characters $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ and an X-tree $\mathcal{T}$, the parsimony score of $\mathcal{C}$ on $\mathcal{T}$, $l_{\mathcal{T}}(\mathcal{C})$ is equal to the sum of $l_{\mathcal{T}}(\chi_i)$ for $1 \leq i \leq k$. An X-tree $\mathcal{T}$ that minimizes $l_{\mathcal{T}}(\mathcal{C})$ is said to be a *maximum parsimony tree*, and the minimum value of $l_{\mathcal{T}}(\mathcal{C})$, written as $l(\mathcal{C})$, is said to be the *maximum parsimony score*. A sequence of characters $\mathcal{C}$ is said to be *compatible* if and only if there is some X-tree $\mathcal{T}$ on which every character is convex.

The parsimony score of two characters can be used to calculate $i(\chi_1, \chi_2)$, the *refined*

incompatibility score ([12]), defined as

$$i(\chi_1, \chi_2) = l(\chi_1, \chi_2) - |\chi_1| - |\chi_2| + 2.$$

3. Subdivision Formulation of Parsimony

In this section we reformulate the maximum parsimony criterion in terms of minimal convex refinements, or minimal subdivisions.

Lemma 3.1. *Let $\chi\colon X \to C$ be a multi-state character and $\mathcal{T}$ a phylogenetic tree. Then there exists a refinement χ' of χ such that $|\chi'| = l_{\mathcal{T}}(\chi) + 1$ and χ' is convex on $\mathcal{T}$.*

Proof. Let $\bar{\chi}$ be a minimal extension of χ to $\mathcal{T}$. Removing the $l_{\mathcal{T}}(\chi)$ edges in $\mathrm{Ch}(\bar{\chi}, \mathcal{T})$ gives $l_{\mathcal{T}}(\chi) + 1$ connected components of T on which $\bar{\chi}$ is constant. As $\bar{\chi}$ is minimal, each component must contain at least one leaf. Define a new character χ' such that the blocks of χ' are in correspondence with the subsets of taxa that label the leaves of each connected component. Then χ' has $l_{\mathcal{T}}(\chi) + 1$ blocks, is convex on $\mathcal{T}$ and if $\chi'(x) = \chi'(y)$ then $\chi(x) = \chi(y)$, so χ' refines χ. ∎

Lemma 3.2. *Let χ be a multi-state character on X and $\mathcal{T}$ a phylogenetic tree. Let χ' be any character that is convex on $\mathcal{T}$ and refines χ. Then $l_{\mathcal{T}}(\chi) \leq |\chi'| - 1$.*

Proof. Let $\bar{\chi'}$ be a minimal extension of χ' to $\mathcal{T}$. Since χ' is convex, removing the edges of $\mathrm{Ch}\big(\bar{\chi'}, \mathcal{T}\big)$ gives $|\chi'|$ connected components that each contains at least one leaf. Define an extension $\bar{\chi}$ of χ to $\mathcal{T}$ by $\bar{\chi}(v) = \chi(l)$ where v is any vertex and l is any leaf in the component that contains v. This extension is well-defined since $\bar{\chi'}$ and hence χ is constant on the leaves of each component. Since $\bar{\chi}$ is constant on every component we have that $\mathrm{ch}(\bar{\chi}, \mathcal{T}) \leq |\chi'| - 1$ and so $l_{\mathcal{T}}(\chi) \leq |\chi'| - 1$. ∎

Theorem 3.3. *Let $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ be a sequence of k characters on X and let $l(\mathcal{C})$ denote the maximum parsimony score. Let B denote the minimum of $\sum_{i=1}^{k} |\chi_i'|$ over all characters $\chi_1', \ldots, \chi_k'$ that refine $\chi_1, \ldots, \chi_k$ respectively and are convex over some tree $\mathcal{T}$. Then $l(\mathcal{C}) = B - k$.*

Proof. Let $\mathcal{T}$ be a maximum parsimony phylogenetic X-tree for $\mathcal{C}$ (note that we may assume $\mathcal{T}$ be a phylogenetic X-tree since such tree can be readily obtained from a non-phylogenetic X-tree). Then by Lemma 3.1 there exist refinements $\chi_1' \cdots \chi_k'$ of $\chi_1 \cdots \chi_k$ that are convex on $\mathcal{T}$ such that $|\chi_i'| = l_{\mathcal{T}}(\chi_i') + 1$. So

$$B \leq l(\mathcal{C}) + k.$$

On the other hand, let $\chi_1' \cdots \chi_k'$ be any characters that refine $\chi_1 \cdots \chi_k$, which are convex on some phylogenetic X-tree $\mathcal{T}$ and satisfy $B = \sum_{i=1}^{k} |\chi_i'|$. Then by Lemma 3.2 $l_{\mathcal{T}}(\chi_i) \leq |\chi_i'| - 1$,

$$l(\mathcal{C}) \leq \sum_{i=1}^{k} l_{\mathcal{T}}(\chi_i) \leq B - k.$$ ∎

Theorem 3.3 can be reformulated in terms of character subdivisions, noting that each subdivision increases the number of blocks by one and that if χ' refines χ then χ' can be obtained from χ through a series of subdivisions. Hence we have

Corollary 3.4. *Let $\mathcal{C} = (\chi_1, \ldots, \chi_k)$ be a sequence of k characters on X. Then the parsimony score is equal to $\sum_i (|\chi_i| - 1)$ plus the number of subdivisions required to transform $\mathcal{C}$ into a sequence of compatible characters.*

In other words, the parsimony score for a sequence of characters equals the minimum number of subdivision required for those characters to have a *perfect phylogeny*, in the technical sense (e.g., [2]).

4. Two Characters-Intersection Graph Approach

We now turn our attention to the problem of computing parsimony scores for pairs of characters. For this, we draw on connections between characters and intersection graphs ([9]). The *intersection graph* for two characters χ_1 and χ_2 has one vertex for every block of χ_1 and χ_2 and an edge between vertices corresponding to blocks that have a non-empty intersection ([16]). We denote this graph by $\Gamma(\chi_1, \chi_2)$. Clearly, $\Gamma(\chi_1, \chi_2)$ is bipartite. The theorem we need can be stated as (from [9]):

Theorem 4.1. ([9]) *Two characters χ_1 and χ_2 on X are compatible if and only if $\Gamma(\chi_1, \chi_2)$ is acyclic.*

Theorem 4.2 can be viewed as a generalisation of Theorem 4.1.

Theorem 4.2. *Let χ_1 and χ_2 be two multi-state characters and $\Gamma(\chi_1, \chi_2) = (V, E)$ the intersection graph for the two characters. Then the maximum parsimony tree for χ_1 and χ_2 has score $|E| + K - 2$, where K is the number of components in $\Gamma(\chi_1, \chi_2)$.*

Proof. Let χ_1' and χ_2' be refinements of χ_1 and χ_2 respectively that are convex on some tree and let $\Gamma(\chi_1', \chi_2') = (V', E')$ be the corresponding intersection graph. Let K' be the number of components of (V', E'). Note that $|V'| \geq |V|$, $|E'| \geq |E|$, and $K' \geq K$ since refining a character cannot decrease any of these quantities. As (V', E') is acyclic we have $|V'| = |E'| + K'$. Hence $|V'| \geq |E| + K$ and, by Theorems 3.3 and 4.1 the maximum parsimony score is at least $|E| + K - 2$.

To show that this minimum can be achieved, it is sufficient to show that if $|E| + K - |V| > 0$, then one of the two characters can be subdivided so that $|V|$ increases by 1 with K and $|E|$ constant. Repeated subdivisions will then achieve the desired minimum.

If $|E| + K - |V| > 0$, then (V, E) contains a cycle. Let $\{w, u\}$ be any edge lying on the cycle, where w corresponds to a block B_1 of χ_1 and u corresponds to a block B_2 of χ_2. As w lies on a cycle of $\Gamma(\chi_1, \chi_2)$ we have that $B_1 - B_2$ is non-empty. Subdivide B_1 into two blocks $B_1 \cap B_2$ and $B_1 - B_2$. The effect on $\Gamma(\chi_1, \chi_2)$ is to replace w by two vertices w_1 and w_2 so that there is an edge $\{w_1, u\}$ and if $\{w, y\}$ is any edge in the old graph, where $y \neq u$, then $\{w_2, y\}$ is an edge in the new graph. The number of edges has not increased. Furthermore, there is a path from u to w_2 along the other edges in the cycle and hence a path from w_1 to w_2. This implies the number of components has not increased either. Therefore we have found a subdivision that increases $|V|$ by 1 but leaves the number of edges and the number of components constant. Repeating this procedure gives a pair of characters χ_1' and χ_2' with $\Gamma(\chi_1', \chi_2') = (V', E')$ and K' components where $|E'| + K' - |V'| = 0$ with Γ acyclic. By Theorems 3.3 and 4.1 the parsimony score for the pair of characters is then $|V'| - 2$ or $|E| + K - 2$. ∎

Note that the linear time calculation for the parsimony score follows from the fact that the intersection graph can be constructed in $O(n)$ time and a depth first search to count the number of components in the graph takes $O(n)$ time, where n is the number of taxa (i.e., $|X| = n$). Interestingly, up to this point, the determination of compatibility of two multi-state characters has implicitly been described as a breadth first search for a cycle within an intersection graph ([14]).

Note that using the framework of Theorem 4.2, the refined incompatibility score for two characters is equal to $|E| + K - |V|$ since $|\chi_1| + |\chi_2| = |V|$. Another result that arises from Theorem 4.2 concerns the upper bound on the maximum parsimony score for two characters.

Corollary 4.3. *Let χ_1 and χ_2 be any two characters on X with $|\chi_1| = r_1$ and $|\chi_2| = r_2$. Then the maximum parsimony score for χ_1 and χ_2 is bounded above by $r_1 r_2 - 1$.*

Proof. By Theorem 4.2 the maximum parsimony score is equal to $|E| + K - 2$ where $|E|$ denotes the number of edges and K denotes the number of components in $\Gamma(\chi_1, \chi_2)$. If $K = 1$ it is easily seen that $r_1 r_2 - 1$ is an upper bound since $\Gamma(\chi_1, \chi_2)$ is a bipartite graph with r_1 and r_2 vertices in each part. Note that adding an edge between any two components cannot decrease the parsimony score. Hence, it is sufficient to consider the case of one component since any upper bound for the many component case is less than or equal to the upper bound for the one component case. ∎

The upper bound in Corollary 4.3 is tight. Set

$$X = \{x_{ij} \colon 1 \le i \le r_1, 1 \le j \le r_2\},$$

and let χ_1 be the character taking x_{ij} to i, χ_2 be the character taking x_{ij} to j, for all i, j. Then $\Gamma(\chi_1, \chi_2)$ has one component and $r_1 r_2$ edges, so that the pair of characters has parsimony score $r_1 r_2 - 1$.

5. Two Characters-Spanning Tree Approach

We now explore the relationship between parsimony trees for two characters and minimum spanning trees, similar to ideas presented in Proposition 5.4.1 of a recent book [16]. A crucial distinction is that Proposition 5.4.1 is stated for a general metric space setting of parsimony (see [16] for details). The following lemma and theorem implicitly assume the discrete metric space setting for parsimony.

Lemma 5.1. *Let $\mathcal{T} = (T, \phi)$ be a maximum parsimony X-tree for two characters χ_1 and χ_2. Then $\mathcal{T}$ can be transformed by a series of edge contractions and rearrangements into a new maximum parsimony X-tree $\mathcal{T}' = (T', \phi')$ such that for every $v \in V(T')$, $\exists x \in X$ where $\phi'(x) = v$.*

Proof. Let $\bar{\chi}_1$ and $\bar{\chi}_2$ be two minimal extensions of χ_1 and χ_2 to $\mathcal{T}$, respectively. First create a new underlying tree T_0 by contracting every edge in $E(T) - (\mathrm{Ch}(\bar{\chi}_1, \mathcal{T}) \cup \mathrm{Ch}(\bar{\chi}_2, \mathcal{T}))$. Let $\mathcal{T}_0 = (T_0, \phi_0)$ be the corresponding X-tree formed by T_0 and ϕ. Note that $l_{\mathcal{T}_0}(\chi_1, \chi_2) = l_{\mathcal{T}}(\chi_1, \chi_2)$ and that the minimal extensions for $\mathcal{T}$ can be mapped into minimal extensions for $\mathcal{T}_0$ as well.

Next, let $v \in V(T_0)$ be any vertex such that $\phi_0(x) \neq v$, $\forall x \in X$. Partition the set of adjacent vertices of v, N_v into three sets: $N_1 = \{u : \bar{\chi}_1(u) = \bar{\chi}_1(v)\}$, $N_2 = \{u : \bar{\chi}_2(u) = \bar{\chi}_2(v)\}$, and $N_3 = N_v - (N_1 \cup N_2)$. Note that N_1 and N_2 are disjoint. Remove v and its $|N_1| + |N_2| + |N_3|$ incident edges. Connect the vertices within each set N_1, N_2, N_3 to give three chains. Finally, let $a_1 \in N_1$, $a_2 \in N_2$, and $a_3 \in N_3$. Create two new edges (a_1, a_2) and (a_1, a_3) thereby creating a new underlying tree T' and corresponding X-tree $\mathcal{T}'$. Note that $\mathrm{ch}(\bar{\chi}_1, \mathcal{T}') \leq \mathrm{ch}(\bar{\chi}_1, \mathcal{T}_0) = \mathrm{ch}(\bar{\chi}_1, \mathcal{T})$ and $\mathrm{ch}(\bar{\chi}_2, \mathcal{T}') \leq \mathrm{ch}(\bar{\chi}_2, \mathcal{T}_0) = \mathrm{ch}(\bar{\chi}_2, \mathcal{T})$. Repeating this procedure for any such v completes the proof. ∎

Note that the series of rearrangements and contractions described in the previous lemma are not unique.

Theorem 5.2. *Let χ_1 and χ_2 be two characters defined on X. Let G be the complete graph on X with edges weights $w(x_1, x_2)$ defined as the Hamming distance between $(\chi_1(x_1), \chi_2(x_1))$ and $(\chi_1(x_2), \chi_2(x_2))$. Then any minimum weight spanning tree of G corresponds to a maximum parsimony X-tree for χ_1 and χ_2.*

Proof. Let T^* be the induced X-tree corresponding to a minimum weight spanning tree of G. Clearly $l_T(\chi_1, \chi_2) \leq l_{T^*}(\chi_1, \chi_2)$, where T is a maximum parsimony X-tree of χ_1 and χ_2. By applying the previous lemma, it is easy to see that T can be transformed into a tree that corresponds to a spanning tree of G showing that $l_{T^*}(\chi_1, \chi_2) \leq l_T(\chi_1, \chi_2)$ thereby completing the proof. ∎

6. Author's Note

Subsequent to the submission of the present article a different algorithm for the two character case (similar to the spanning tree approach) was published independently by Althaus and Naujoks ([1]).

Acknowledgments. The authors would like to thank the two anonymous reviewers for a number of helpful suggestions on the manuscript. T.C. Bruen acknowledges the support of a McGill Major scholarship.

References

1. E. Althaus and R. Naujoks, Computing Steiner minimum trees in Hamming metric, In: Proceedings of the Seventeenth Annual ACM-SIAM Symposium on Discrete Algorithm, ACM, New York, (2006) 172–181.
2. D. Gusfield, Efficient algorithms for inferring evolutionary trees, Networks **21** (1) (1991) 19–28.
3. B.R. Holland, K.T. Huber, D. Penny, and V. Moulton, The minmax squeeze: guaranteeing a minimal tree for population data, Mol. Biol. Evol. **22** (2) (2005) 235–242.
4. J.P. Huelsenbeck and F. Ronquist, MRBAYES: bayesian inference of phylogenetic trees, Bioinform. **17** (8) (2001) 754–755.
5. M. Steel and D. Penny, Maximum parsimony and the phylogenetic information in multi-state characters, In: Parsimony, Phylogeny and Genomics, V. Albert Ed., Oxford University Press, (2005) pp. 163–178.
6. L.R. Foulds and R.L. Graham, The steiner problem in phylogeny is NP-complete, Adv. Appl. Math. **3** (1) (1982) 43–49.

7. R. Karp, Reducibility among combinatorial problems, In: Complexity of Computer Computations, R.E. Miller and J.W. Thatcher Eds., Plenum Press, New York, (1972) pp. 85–104.
8. M.R. Garey and D.S. Johnson, Computers and Intractibility, W.H. Freeman & Co., 1979.
9. G.F. Estabrook and F.R. McMorris, When are two qualitative taxonomic characters compatible?, J. Math. Biol. **4** (1977) 195–200.
10. D.F. Robinson and L.R. Foulds, Comparison of phylogenetic trees, Math. Biosci. **53** (1981) 131–147.
11. J. Felsenstein, Evolutionary trees from DNA sequences: a maximum likelihood approach, J. Mol. Evol. **17** (6) (1981) 368–376.
12. T. Bruen, H. Philippe, and D. Bryant, A simple and robust statistical test to detect the presence of recombination, Genetics **172** (2006) 2665–2681.
13. J.H. Camin and R.R. Sokal, A method for deducing branching sequences in phylogeny, Evolution **19** (3) (1965) 311–326.
14. G.F. Estabrook and L. Landrum, A simple test for the possible simultaneous evolutionary divergence of two amino acid positions, Taxon **24** (5/6) (1975) 609–613.
15. D. Penny and M. Hendy, Estimating the reliability of evolutionary trees, Mol. Biol. Evol. **3** (5) (1986) 403–417.
16. C. Semple and M. Steel, Phylogenetics, Oxford University Press, Oxford, 2003.